Enhanced Event Extraction from Text via Error-Driven Aggregation Methodologies

Tracy D. Lemmond
(925) 422-0219
lemmond1@llnl.gov

Knowledge discovery systems are designed to construct massive data repositories via text and information extraction methodologies, and then infer knowledge from the ingested data, allowing analysts to "connect the dots." The extraction of relational information (such as triples and events) and related entities (such as people and organizations) generally forms the basis for data ingestion. Unfortunately, these systems are particularly vulnerable to errors introduced during the ingestion process, often resulting in misleading or unreliable analysis. Though state-of-the-art extraction tools achieve insufficient accuracy rates for practical use, not all extractors are prone to the same types of error. This suggests that substantial improvements may be achieved via appropriate combinations of existing extraction tools, provided their behavior can be accurately characterized and quantified.

Our research is addressing this problem via the aggregation of extraction tools based on a general inferential framework that exploits their strengths and mitigates their weaknesses.

Project Goals

The objective of this effort is to develop a significantly improved entity/event extraction system that will enable 1) greater insight into the downstream effects of extraction errors; 2) more accurate automatic text extraction; 3) better estimates of uncertainty in extracted data; 4) effective use of investments by the Natural Language Processing community; and 5) rapid incorporation of future advancements in extraction technologies.

An extensive analysis of the error processes of individual extractors will yield insights into their synergistic and conflicting behaviors. These insights will be leveraged to configure a collection of base extractors, via a general inferential framework, into an aggregate meta-extractor with substantially improved extraction performance (Figs. 1 and 2).

Relevance to LLNL Mission

Nonproliferation, counterterrorism, and other national security missions rely on the acquisition of knowledge that is buried in unstructured text documents too numerous to be manually

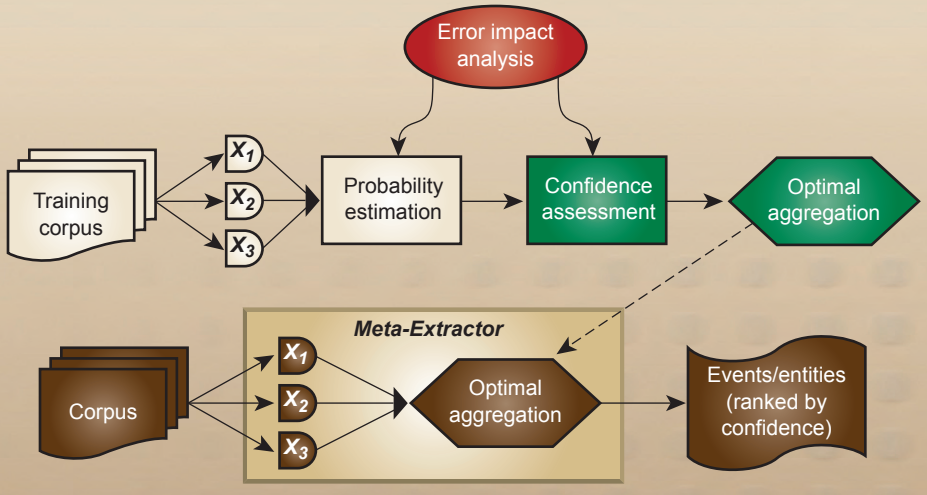

Figure 1. Schematic of the meta-extraction system.

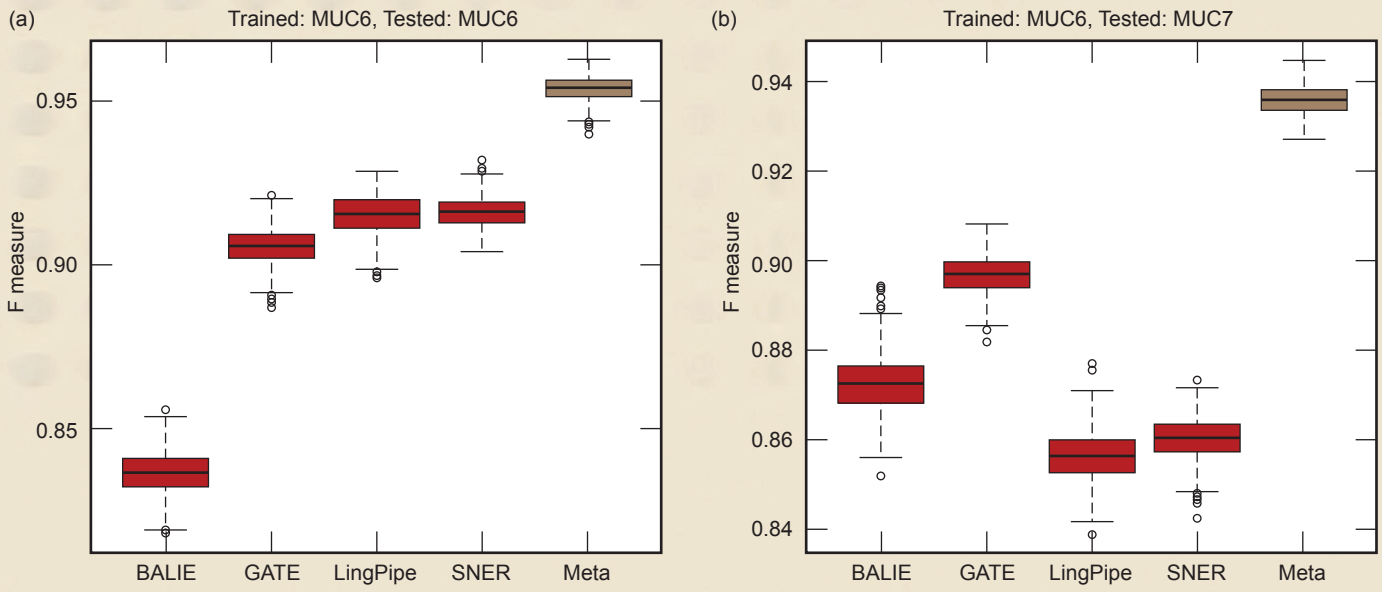

Figure 2. **Box plots showing bootstrapped samples of the weighted mean of F measure. The data come from MUC (Message Understanding Conferences) 6 and 7; [(a) and (b), respectively]. The meta-extractor shows statistically significant improvement over the base extractors.**

processed. Systems are under development by LLNL and its customers that must automatically extract critical information from these sources. To enable effective knowledge discovery, however, extraction error rates must be driven down. Probabilistic aggregation of extractors is a promising and innovative approach to accomplishing this goal. This effort directly supports the Engineering Systems for Knowledge and Inference (ESKI) focus area, the Text to Inference R&D area, and the Cyber, Space, and Intelligence strategic mission thrust in the LLNL five-year strategic roadmap. Successful completion of this research will provide highly valued and unprecedented inference and decision-making capabilities to internal programs, such as IOAP and CAPS, and to external customers such as DHS, DoD, and the IC.

FY2009 Accomplishments and Results

Insights gained in event extraction error analyses performed in FY2008 motivated a graduated approach to triple/event aggregation that is founded upon the aggregation of extracted entities. To this end, we have developed a novel aggregation methodology (the "meta-extractor") focused on entity extraction that can be generalized to multiscale triples (for example, simple events) and more complex event aggregation solutions. This methodology couples an innovative pattern-based approach with probabilistic methods to characterize the marginal and joint behaviors of entity extraction tools and aggregate their output. Empirical studies have shown that the developed aggregation methodologies achieve up to 120% improvement in performance over the best extraction tool included in the study. Moreover, these methodologies have been shown capable of determining the true entities when *all* of the extraction tools fail, a significant achievement in natural language processing.

Related References

1. Chen, M., Q. Shao, and J. Ibrahim, Monte Carlo Methods in Bayesian Computation, Springer, 2000.
2. Efron, B., and R. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, 1993.
3. Kohavi, R., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence*, **2**, 12, pp. 1137–1143, 1995.

FY2010 Proposed Work

In FY2010, we will extend and enhance the entity aggregation methodologies developed in FY2009 to more effectively leverage other information, such as entity type. In addition, we will continue to develop and generalize these methods to address the more general triple aggregation task. This unprecedented work will synergistically leverage the insights gained from the event error analyses performed in FY2008 and the entity aggregation research performed in FY2009 to produce a state-of-the-art extensible methodology for triple aggregation. If successful, our research will represent a significant breakthrough in natural language processing and knowledge discovery.